Final capstone
Twitter Sentiment Analysis

Presented by: Mohammad Nosrati

Data provided by: Kaggle.org
Thinkful Data Science Bootcamp

Problem

- The problem in sentiment analysis is classifying the polarity of a given tweet or text regarding any service, product, or company, whether the expressed opinion in the tweet is positive or negative.

- How to analyze millions of tweets and classify them as positive or negative to improve marketing or the quality of the service.

Model and Analysis

- **Sentiment Analysis** is a very frequent term within text classification and is essentially to use natural language processing (quite often referred simply as NLP)+ machine learning to interpret and classify emotions in text information.

- Sentiment analysis of the tweets for this project has been done with three models such as basic logistic regression, LSTM, and AutoNLP.
- To pick the best performing model the performance metrics of each model with train and test datasets has been used.

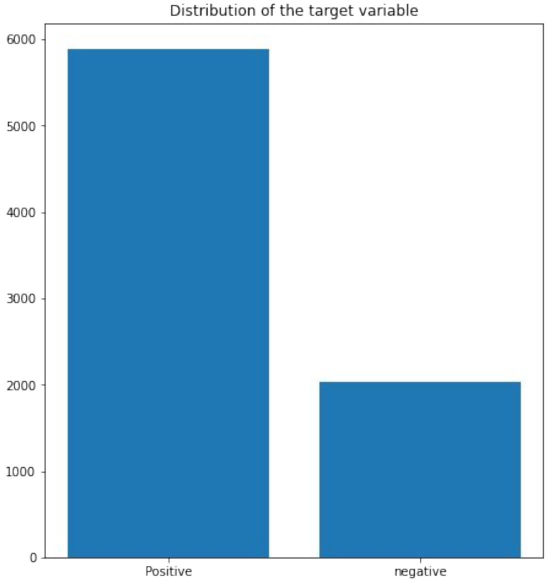
Logistic regression model

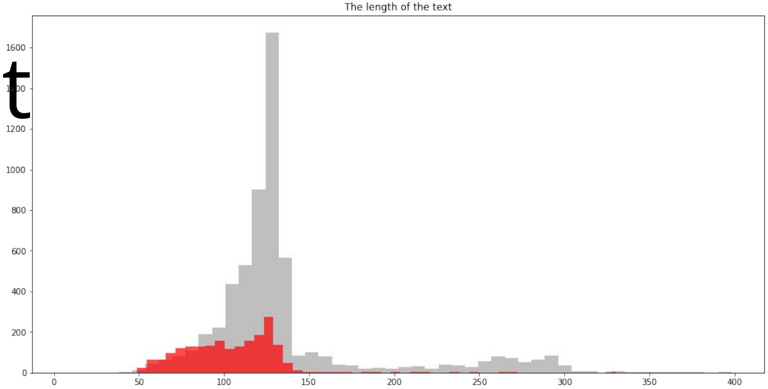
- logistic regression. logistic regression is one of the most important analytic tools in the social and natural sciences. In natural language processing, logistic regression is the baseline supervised machine learning algorithm for classification and has a very close relationship with neural networks.

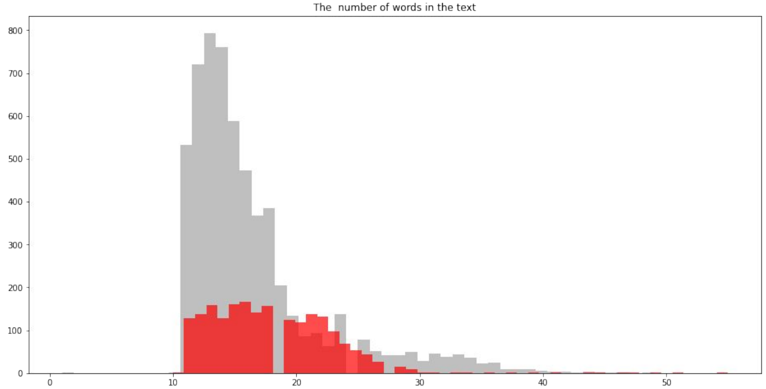
- To perform a logistic regression after tokenizing, cleaning/preprocessing text dataset such as removing punctuations, emojis, stopwords, HTML tags, and lemmatization. The text has been vectorized by using Bag of words and TF-IDF methods and then run the model.

Text dataset

```
Class  -> Counts -> Percent
          0:     4713   ->   74.4%
          1:     1623   ->   25.6%
```

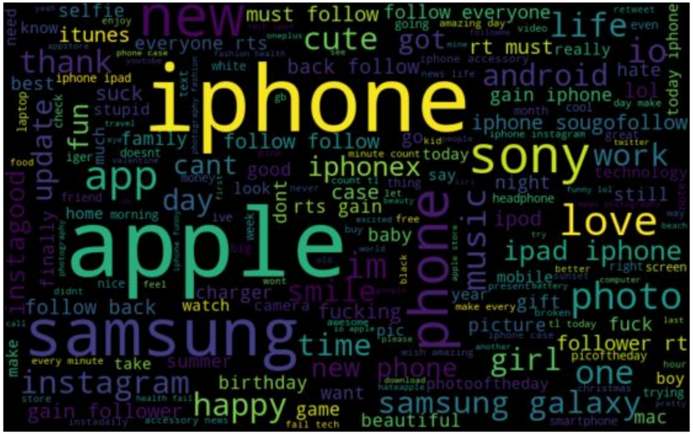

Distribution of the target variable


The length of the text

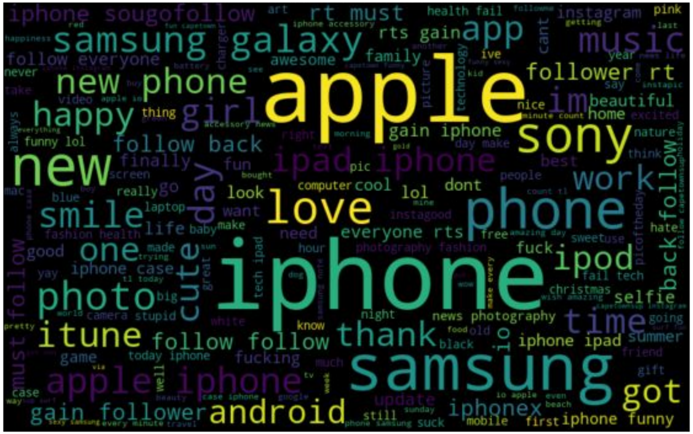

The number of words in the text

Train and test datasets

Most Common words in column tweet lemmatized



Most Common words in column tweet lemmatized



LSTM

- Long Short-Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies.

- LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

- The LSTM model performed after cleaning/ preprocessing, encoding and word embedding the text dataset.

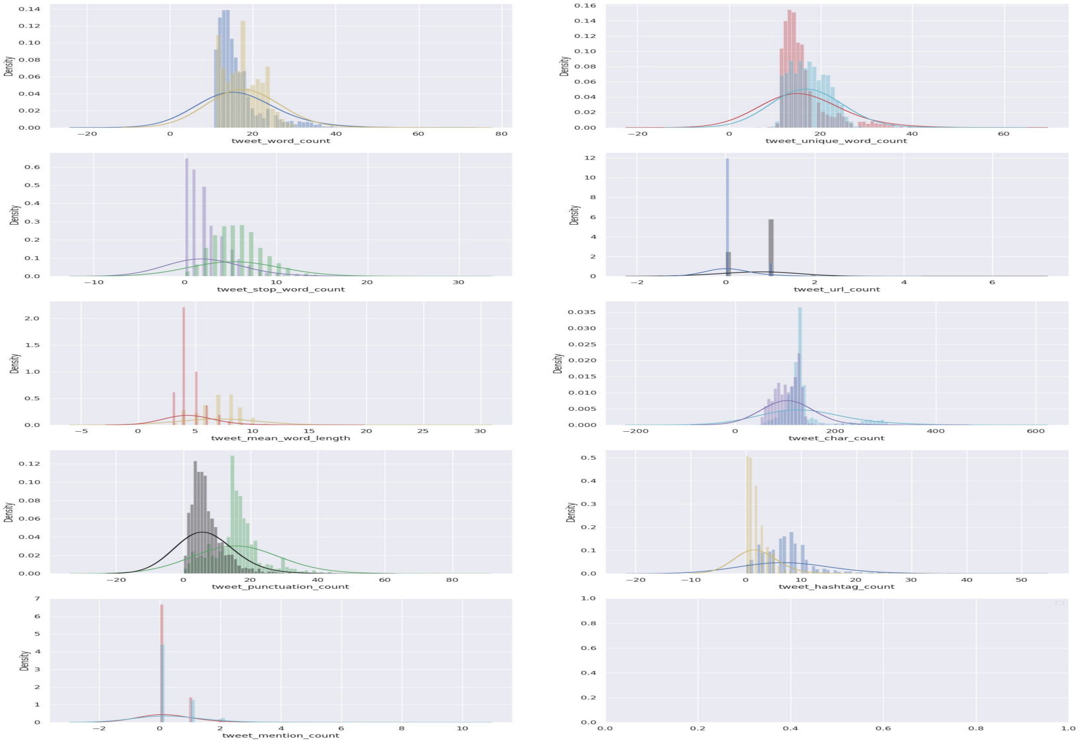
AutoNLP

- AutoNLP: Auto training and fast deployment for state-of-the-art NLP models

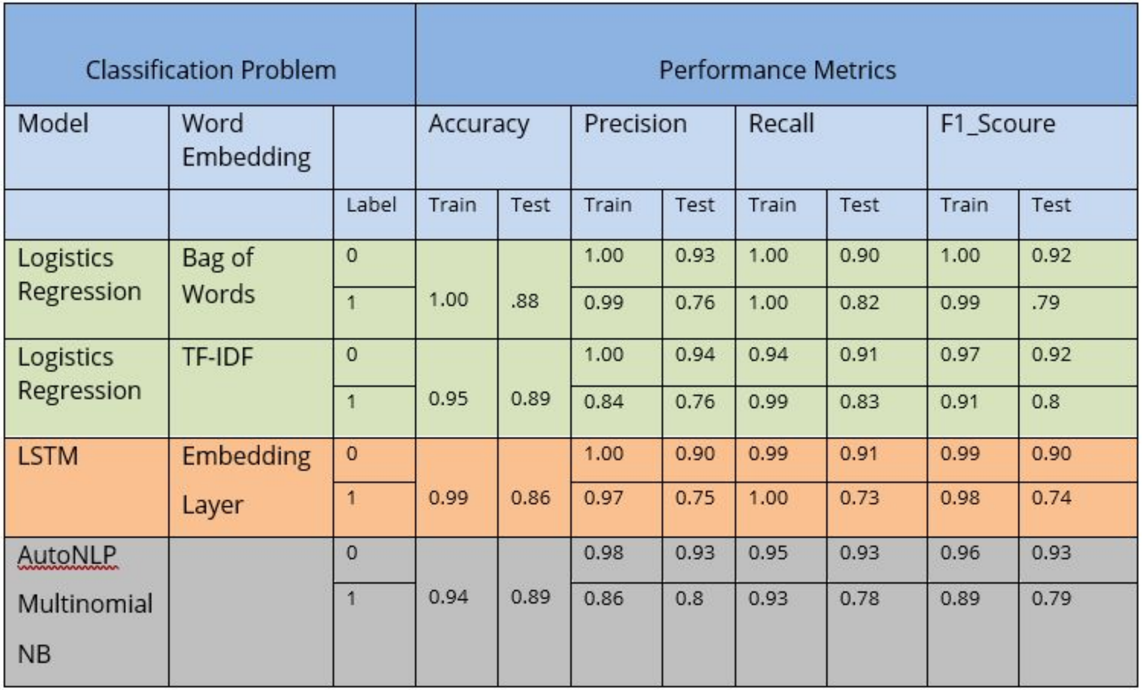
- AutoNLP is an automatic way to train, evaluate and deploy state-of-the-art NLP models for different tasks. Using AutoNLP, you can leave all the worries of selecting the best model, fine-tuning the model or even deploying the models and focus on the broader picture for your project/business.

- After installing the packages and uploading the dataset it was ready to run the model.

AutoNLP


Density
tweet_word_count
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
−20
0
20
40
60
80
tweet_unique_word_count
0.16
tweet_stop_word_count
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−10
10
30
tweet_url_count
12
8
6
4
2
−2
tweet_mean_word_length
0.5
1.0
1.5
2.0
−5
5
15
25
tweet_char_count
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
−200
200
400
600
tweet_punctuation_count
tweet_hashtag_count
50
tweet_mention_count
7
3
1
0.8
0.6
0.4
0.2
1.0

Model Evaluation



Classification Problem			Performance Metrics							
Model	Word Embedding		Accuracy		Precision		Recall		F1_Scoure	
		Label	Train	Test	Train	Test	Train	Test	Train	Test
Logistics Regression	Bag of Words	0	1.00	.88	1.00	0.93	1.00	0.90	1.00	0.92
		1			0.99	0.76	1.00	0.82	0.99	.79
Logistics Regression	TF-IDF	0	0.95	0.89	1.00	0.94	0.94	0.91	0.97	0.92
		1			0.84	0.76	0.99	0.83	0.91	0.8
LSTM	Embedding Layer	0	0.99	0.86	1.00	0.90	0.99	0.91	0.99	0.90
		1			0.97	0.75	1.00	0.73	0.98	0.74
AutoNLP Multinomial NB		0	0.94	0.89	0.98	0.93	0.95	0.93	0.96	0.93
		1			0.86	0.8	0.93	0.78	0.89	0.79

0-Positive, 1-Negative

Recommendation

The dataset is imbalanced and accuracy in this case will be misleading.
In case of having false positive rate and false negative rates equally important f1 scour would be better metric to compare the models.

Since the gap between the train and test set metrics is big there is an overfitting problem

Among all four models the difference between train and test metrics is less in AutoNLP. It is safe to say that there is no overfitting problem or it is minimal.

Conclusion

The AutoNLP model will be the best option to Classify tweets for Sentiment Analysis.